회귀분석팀

6팀

심은주
진수정
문병철
이수정
임주은

INDEX

1

회귀분석이란?

- **회귀분석의 정의**

둘 또는 그 이상의 변수들 간의 인과관계를 파악하고, 이를 통해 특정 변수의 값을 다른 변수들을 이용하여 설명하고 예측하는 분석

- **회귀식**

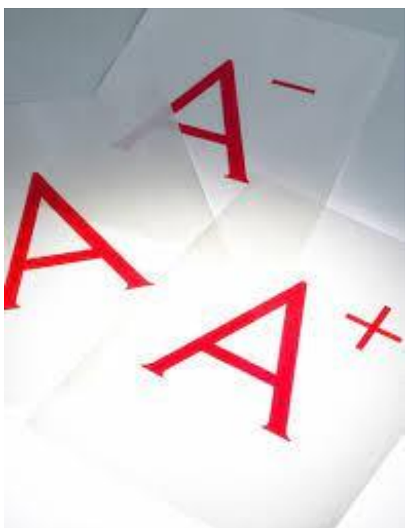
$$Y = f(X_1, X_2, ..., X_p) + \varepsilon$$

- Y : 반응변수(Response Variable), 종속변수(Dependent Variable)
- X : 설명변수(Explanatory Variable), 예측변수(Predictor Variable)
- ε : 오차항(random error), 모형이 데이터를 정확하게 적합하지 못하는 정도
- f : 독립변수들 간의 관계

- **회귀모델링 과정**

예시

학점과 통학거리는 관련이 있을까??



< 학점과 통학거리는 어떤 관계가 있을까?>

1. 문제 정의

"학점을 가장 잘 표현 할 수 있는 **변수**들은 무엇이 있을까?"

2. 변수 선택

X1~X3 : 통학 거리, SNS 사용시간, 듣는 학점 수

3. 데이터 수집 및 전처리

학점, 집주소와 학교 사이의 거리 계산, 시간표, 휴대폰 SNS 사용 시간

4. 모형 설정과 적합

선형 vs 비선형 / 단순회귀 vs 다중 회귀 / 일변량 vs 다변량 등을 고려

학점 = a*통학거리 + b*SNS 사용시간 + c*듣는 학점 수

5. 모형 평가

모형이 회귀 가정을 만족하는가?

2주차에서 만나요…!

6. 모형 해석

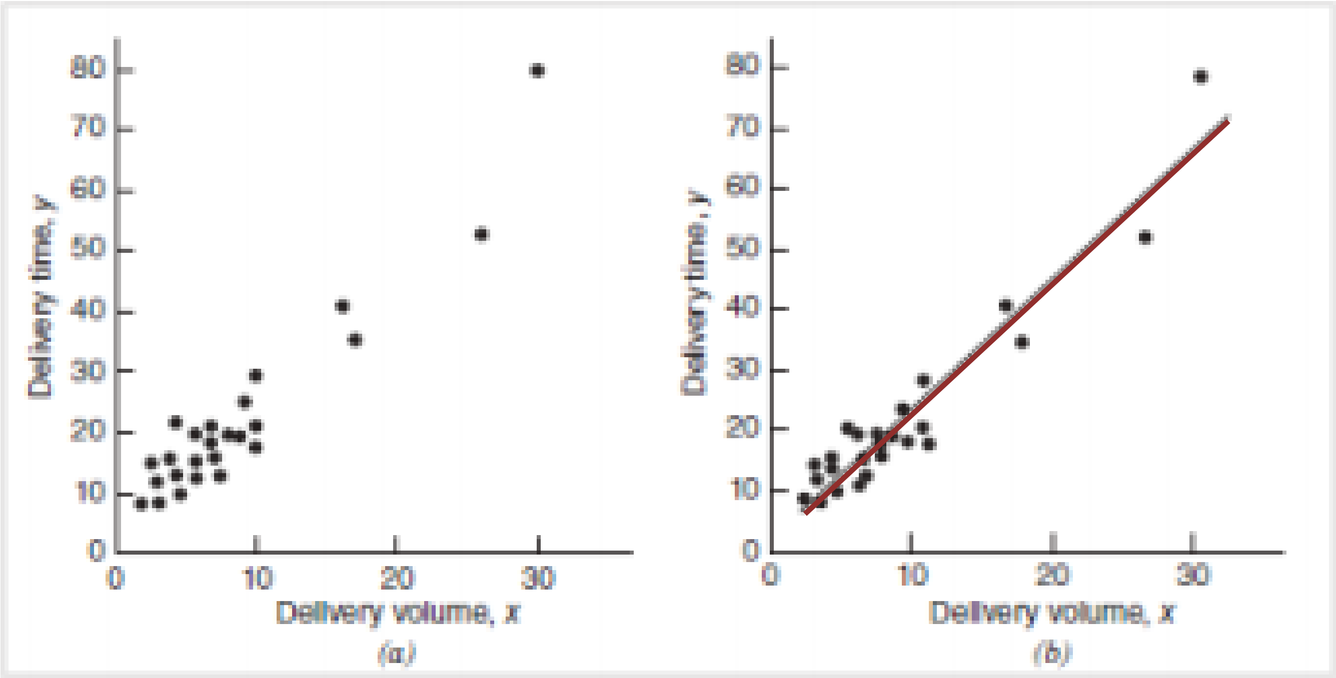
15학점을 수강하고, 하루 평균 SNS 3시간 사용하며,현재 주거지에서 통학 할 때 학점은 평균적으로 4.0정도 일 것이다.

2

단순선형회귀

- **단순선형회귀**

하나의 X변수와 Y변수의 관계를 가장 잘 표현 할 수 있는 **직선**을 찾는 것



- 단순선형회귀식

Population regression model (모집단의 관점)

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Sample regression model (관측치의 관점)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

ε_i : i 번째 관찰값에 의한 랜덤 오차 , $\varepsilon_i \sim NID(0, \sigma^2)$

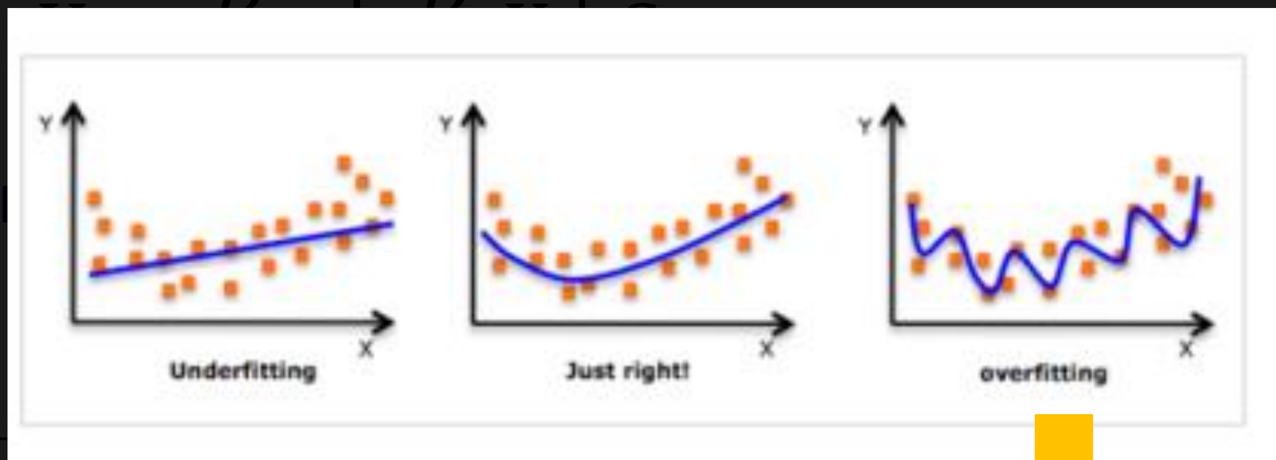
$\beta_0,\ \beta_1$: 회귀계수 또는 우리가 추정해야 할 모수

회귀계수를 잘 추정하는 것이 가장 중요!!

왜 직선인가요?

• 단순선형회귀식

변수의 영향력을 간단하게 모형화 할 수 있기 때문!

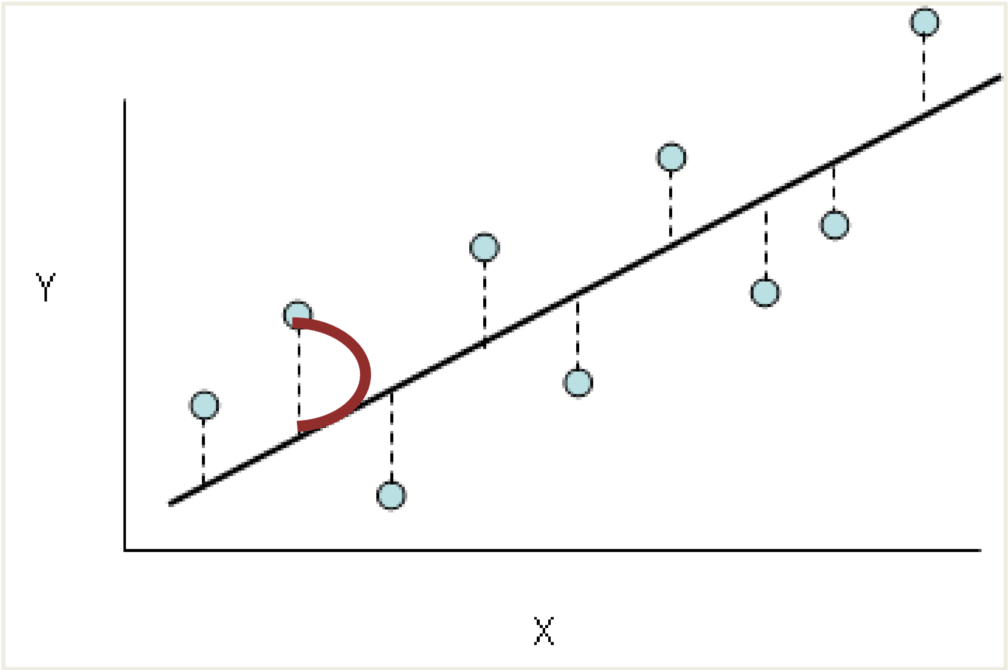


ε_i : i 번째 관찰값에 의한 랜덤 오차 , $\varepsilon_i \sim NID(0, \sigma^2)$

고차근사로 가면, 당장의 데이터는 잘 설명해도
다른 데이터에 모델을 적용했을 때 잘 설명하지 못하는
과적합 문제가 발생한다!

- **모수 추정 - 최소제곱법(Least Square Estimation Method)**

각 점으로부터 구하고자 하는 최적 직선까지의 수직거리의 **제곱합을 최소**로 하는 방법



실제 데이터와 우리가 추정한 값의 오차가 작을 수록 좋은 추정

- **모수 추정 - 최소제곱법(Least Square Estimation Method)**

각 점으로부터 구하고자 하는 최적 직선까지의 수직거리의 **제곱합을 최소**로 하는 방법

오차항의 제곱합을

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} {\varepsilon_i}^2$$

편미분하여 오차를 최소화하는

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1) = 0$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1) x_i = 0$$

- **모수 추정 - 최소제곱법(Least Square Estimation Method)**

각 점으로부터 구하고자 하는 최적 직선까지의 수직거리의 **제곱합을 최소**로 하는 방법

모수를 추정한 뒤

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

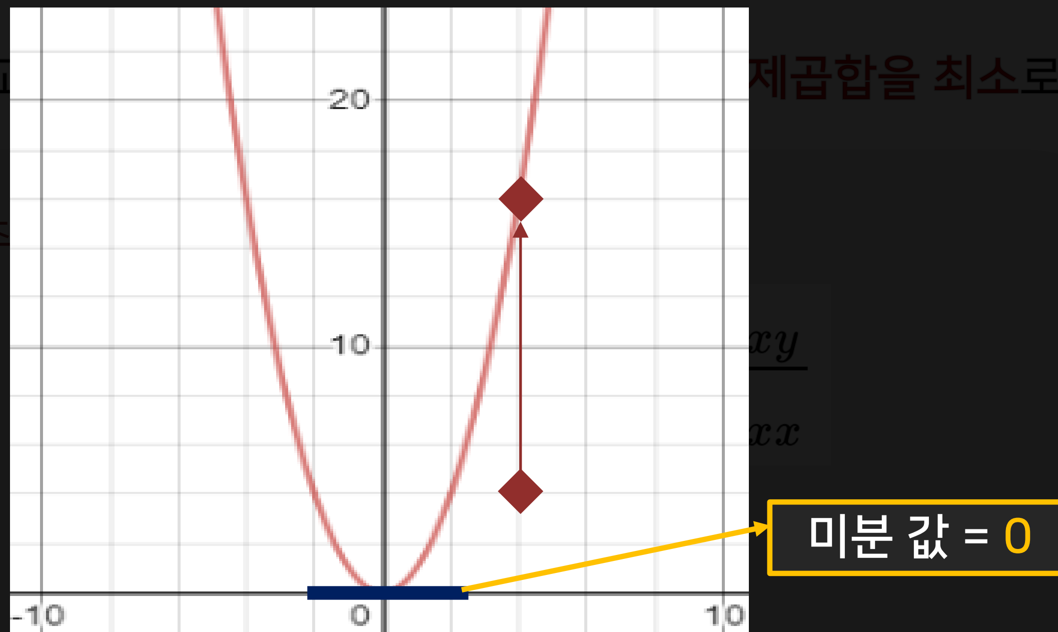
회귀식을 도출

$$\hat{\boldsymbol{y}} = \widehat{\boldsymbol{\beta}}_{\mathbf{0}} + \widehat{\boldsymbol{\beta}}_{\mathbf{1}}\mathbf{x}$$

모수들은 모두 추정치이기 때문에
hat 을 사용!

왜 오차제곱합인가요?

- 모수 추정 - 최소제곱법(Least Square Estimation Method)



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

1) 미분이 편리하고,

2) 오차가 클수록 더 큰 패널티를 부여할 수 있기 때문!

- 모수 추정 - BLUE

  - BLUE(Best Linear Unbiased Estimator)

    : 선형의 불편추정량 중 분산이 가장 작은 추정량

- 오차들의 평균은 0
- 오차들의 분산은 σ^2 으로 동일
- 오차 간에는 자기상관이 없음

* 정규성 조건은 필요하지 않음

세 가지 조건이 충족될 때, 최소제곱추정량은 BLUE!

- MLE vs. LSE

- **MLE (Maximum Likelihood)**

확률적인 방법에 근거해, 데이터가 나올 "가능도"를 최대로 하는 모수를 선택하는 방법

- ML은 분포 가정이 필수적!

$\varepsilon_i \sim N(0, \sigma^2)$ 라는 정규분포 가정이 있다면,

MLE는 LSE와 완전히 동일한 추정량을 가짐

- 적합성(Goodness of fit) 검정
  - 잔차(Residual)란?

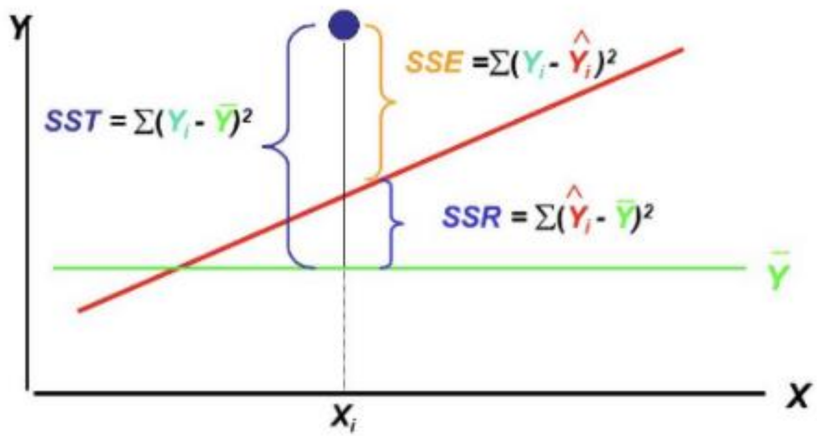
  : 오차의 추정량

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \ , \ \sum e_i = 0$$

모집단일 때는 오차, 표본일 때는 잔차를 쓸 뿐,

결과적으로 큰 차이는 없다!

- 적합성(Goodness of fit) 검정
  - 잔차를 통한 적합성 검정



SST: 종속변수 Y의 총 변동

SSR: 회귀선이 설명하는 변동

SSE: 회귀선이 설명하지 못하는 변동

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

$$\therefore SST = SSR + SSE$$

- 적합성(Goodness of fit) 검정
  - 잔차를 통한 적합성 검정

결정 계수: 총 변동에서 회귀식이 설명하는 부분

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- 범위: $0 \leq R^2 \leq 1$

- 값이 클수록 회귀식으로 설명되는 변동의 비율이 크므로, 1에 가까울수록 좋음

- 유의성 검정

$\varepsilon_i \sim N(0, \sigma^2)$ 라는 정규분포 가정하에 개별 베타 계수에 대한 통계적 검정 가능

귀무가설	H_0: $\beta = 0$
대립가설	H_1: $\beta \neq 0$

귀무가설을 기각하지 못하여도,

X와 Y 사이에 선형적 관계가 없을 뿐

아무 의미가 없는 것은 아님!

3

다중선형회귀

- **다중선형회귀란?**

여러 개의 설명변수 X와 종속변수 Y의 관계를 표현한 식을 찾는 것

단순선형회귀

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

설명변수를 p개로 확장

다중선형회귀

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

- **다중선형회귀란?**

여러 개의 설명변수 X와 종속변수 Y의 관계를 표현한 식을 찾는 것

단순선형회귀

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

설명변수를

p개로 확장

다중선형회귀

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

나머지 X 변수들이 고정되었을 때,
x_1이 1단위 증가하면 y는 평균적으로 β_1만큼 증가함을 의미

- **모수의 추정: 최소제곱법(LSE)**

단순선형회귀와 동일한 방식으로 모수의 추정치 산출 가능

$$S(\beta) = \sum_i (y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi})^2$$

$$\frac{\partial S}{\partial \beta_0} = -2\sum_i \left(y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}\right) = 0$$

$$\vdots$$

$$\frac{\partial S}{\partial \beta_p} = -2\sum_i \left(y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}\right) x_{pi} = 0$$

계산이

매우 복잡하다…!

행렬을 이용하자!

- **모수의 추정: 최소제곱법(LSE)**

$$Y=\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad X=\begin{pmatrix} 1 & \cdots & x_{p1} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{pn} \end{pmatrix} \quad \beta=\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \quad \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

회귀식 $\mathrm{Y}=X\beta+\varepsilon$

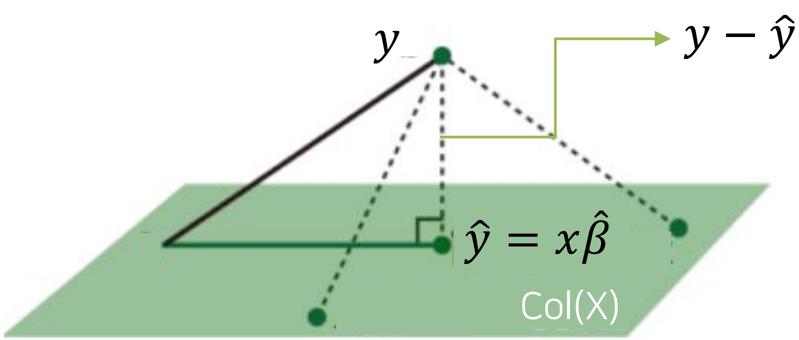
목적함수 $\mathrm{S}(\beta)=\sum_i \varepsilon_i^2=(\mathrm{Y}-X\beta)'(\mathrm{Y}-X\beta)$

Normal Equation $\frac{\partial S}{\partial \beta}=-2X'(\mathrm{Y}-X\hat{\beta})=0$

$$\hat{\beta}^{LSE}=argmin\,S(\beta)=(X'X)^{-1}X'Y$$ when $(X'X)^{-1}$ exists

- **모수의 추정: 최소제곱법(LSE)**

- Normal Equation의 기하학적 해석



$$1 \perp (Y - X\hat{\beta})$$
$$x_1 \perp (Y - X\hat{\beta})$$
$$\vdots$$
$$x_p \perp (Y - X\hat{\beta})$$

$$1'(Y - X\hat{\beta}) = 0$$
$$x_1'(Y - X\hat{\beta}) = 0$$
$$\vdots$$
$$x_p'(Y - X\hat{\beta}) = 0$$

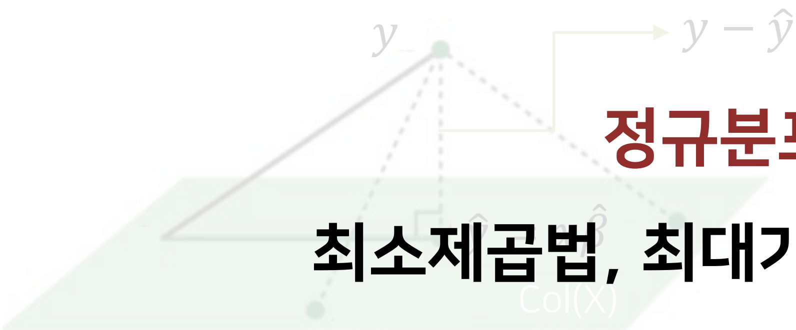
$\hat{y}$ 은 y 를 $Col(X)$ 에 프로젝션 시킨 것이기 때문에

$y-\hat{y}$은 $Col(X)$에 수직

$$X'(Y - X\hat{\beta}) = 0$$

- 모수의 추정: 최소제곱법(LSE)

- Normal Equation의 기하학적 해석



$1 \perp (Y - X\hat{\beta})$
$x_1 \perp (Y - X\hat{\beta})$
⋮
$x_p \perp (Y - X\hat{\beta})$

$1'(Y - X\hat{\beta}) = 0$
$x_1'(Y - X\hat{\beta}) = 0$
⋮
$x_p'(Y - X\hat{\beta}) = 0$

$X'(Y - X\hat{\beta}) = 0$

정규분포 가정 하에서

최소제곱법, 최대가능도추정법 결과는 동일!

$$\hat{\beta}^{LSE} = \hat{\beta}^{MLE}$$

$\therefore\ \hat{\beta}^{LSE} = (X'X)^{-1}X'Y \qquad$ when $(X'X)^{-1}$ exists

- **유의성 검정: 회귀식의 독립변수가 통계적으로 유의미한가?**

1. F-test: 모델 전체에 대한 검정

$$\boldsymbol{H_0}:\ \boldsymbol{\beta_1 = \beta_2 = \cdots = \beta_p = 0}$$

$\boldsymbol{H_1}$: $\boldsymbol{\beta_1, \beta_2, \ldots, \beta_{\mathrm{p}}}$ 중 적어도 하나는 0 이 아니다.

검정통계량

$$F = \frac{\frac{SST - SSE}{p}}{\frac{SSE}{n-p-1}} = \frac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}} = \frac{MSR}{MSE}$$

- 유의성 검정: 회귀식의 **독립변수**가 통계적으로 **유의미**한가?

1. F-test: 모델 전체에 대한 검정

$$\boldsymbol{H_0}: \ \boldsymbol{\beta_1 = \beta_2 = \cdots = \beta_p = 0}$$

$\boldsymbol{H_1}$: $\boldsymbol{\beta_1, \beta_2, \ldots, \beta_\mathrm{p}}$ 중 적어도 하나는 0 이 아니다.

if 기각되지 않는다면?

$$y = \beta_0 + \varepsilon \quad (\because \beta_1 = \ \beta_2 = \cdots = \beta_p = 0)$$

회귀식이 아무런 **의미가 없음**을 의미!

전부 $\bar{y}$로 예측하게 됨…

- 유의성 검정: 회귀식의 독립변수가 통계적으로 유의미한가?

2. Partial F-test

FM(Full Model) : 모든 변수를 사용한 회귀모형

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

RM(Reduced Model) : 일부 회귀계수를 특정한 값으로 둔 축소모형

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_q x_{qi} + \varepsilon_i \quad \text{where q} < \text{p}$$

- 유의성 검정: 회귀식의 독립변수가 통계적으로 유의미한가?

2. Partial F-test

H_0: $\boldsymbol{\beta}_{q+1} = \boldsymbol{\beta}_{q+2} = \cdots = \boldsymbol{\beta}_p = \mathbf{0}$ RM이 적절

H_1: $\boldsymbol{\beta}_{q+1}, \boldsymbol{\beta}_{q+2}, \dots, \boldsymbol{\beta}_p$ 중 적어도 하나는 0이 아니다. FM이 적절

검정통계량

$$F = \frac{\frac{SSR(FM) - SSR(RM)}{p-q}}{\frac{SSE(FM)}{n-p-1}} \sim F_{p-q,n-p-1}$$

- **유의성 검정: 회귀식의 독립변수가 통계적으로 유의미한가?**

2. Partial F-test

H_0: $\boldsymbol{\beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0}$ RM이 적절

H_1: $\boldsymbol{\beta_{q+1}, \beta_{q+2}, \ldots, \beta_p}$ 중 적어도 하나는 0이 아니다. FM이 적절

if $F \geq F_{p-q,n-p-1;\alpha}$

귀무가설 기각

추가된 변수들이 설명력을 유의미하게 증가시키므로, FM이 적절!

- **유의성 검정: 회귀식의 독립변수가 통계적으로 유의미한가?**

2. Partial F-test

H_0: $\boldsymbol{\beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0}$ RM이 적절

H_1: $\boldsymbol{\beta_{q+1}, \beta_{q+2}, \ldots, \beta_p}$ 중 적어도 하나는 0이 아니다. FM이 적절

- 회귀식 전체에 대한 F-test는 Partial F-test의 한 케이스임

$$F = \frac{\frac{SSR(FM) - SSR(RM)}{p}}{\frac{SSE(FM)}{n-p-1}} = \frac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}} = \frac{MSR}{MSE}$$

$$\sum_i (\bar{Y} - \bar{Y})^2 = 0$$

- **유의성 검정: 회귀식의 독립변수가 통계적으로 유의미한가?**

3. t-test

: 개별 회귀계수의 유의성을 검정함

H_0: $\beta_j = 0$

다른 변수들이 적합된 상태에서 x_j는 통계적으로 유의하지 않다

H_1: $\beta_j \neq 0$

다른 변수들이 적합된 상태에서 x_j는 통계적으로 유의하다

검정통계량

$$t_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)} \sim t_{n-p-1}$$

- **유의성 검정: 회귀식의 독립변수가 통계적으로 유의미한가?**

3. t-test

: 개별 회귀계수의 유의성을 검정함

$H_0: \ \beta_j = 0$

다른 변수들이 적합된 상태에서 x_j는 통계적으로 유의하지 않다

$H_1: \ \beta_j \neq 0$

다른 변수들이 적합된 상태에서 x_j는 통계적으로 유의하다

if $|t_j| \geq t_{n-p-1;\alpha/2}$

→ 귀무가설 기각

→ 다른 변수들이 적합된 상태에서, x_j는 통계적으로 유의한 변수

T-test로 변수를 선택하는 것이 가능할까?

T-test는 다른 변수들이 고정된 상태에서
해당 변수의 추가가 유의미한 설명력 증가를 가져오는지 판단하는 것

다른 회귀식을 가정하면
해당 변수의 유의성 바뀔 수 있음

T-test로 변수를 선택하는 것은 위험!

변수 선택법은 3주차에서…

T-test를 하기 전에

회귀식 전체에 대한 F-test를 먼저 확인해야 하는 이유?

- 전체 회귀식에 대한 검정이 더 엄격하기 때문
- F-test를 기각 못 해도 몇몇 T-test는 기각하는 경우가 있을 수 있기 때문

이 경우 F-test는 유의수준을 충족하지만,

T-test의 Type1 error는 유의수준보다 커지기 때문에

T-test 결과를 신뢰할 수 없음

∴ F-test 결과를 먼저 확인해야 함

- 적합성 검정: 모형이 주어진 데이터를 잘 설명하는가?

R^2값을 통해
데이터를 잘 설명하는 모델을 찾을 수 있을까?

No!

R^2의 문제점

변수가 늘면, 항상 값이 증가함

총 변동은 고정되어 있는데, 변수가 추가되면 회귀식으로 설명되는 변동이 조금이라도 증가할 수 밖에 없기 때문

∴ 변수의 개수가 다른 두 회귀모형의 직접적인 비교가 어려움

- 적합성 검정: 모형이 주어진 데이터를 잘 설명하는가?

수정결정계수

$$R_a^2 = \frac{SSP/p}{SST/(n-1)} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

- SSE와 SST를 각각의 자유도로 나누어 계산한 형태

- R_a^2 값이 더 높은 회귀식이 더 좋은 회귀식

- 변수의 개수가 다른 두 회귀식을 비교할 때 사용 가능

- 예시: 고객의 특성(X)과 신용카드 잔액(Y)의 관계를 표현한 회귀식

```
> model1 = lm(Balance ~ Income + Limit + Rating + Cards + Age + Education, data = Credit)
> summary(model1)

Call:
lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
    Education, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max
-227.25 -113.15  -42.06   45.82  542.97

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -477.95809   55.06529  -8.680  < 2e-16 ***
Income        -7.55804    0.38237 -19.766  < 2e-16 ***
Limit          0.12585    0.05304   2.373  0.01813 *
Rating         2.06310    0.79426   2.598  0.00974 **
Cards         11.59156    7.06670   1.640  0.10174
Age           -0.89240    0.47808  -1.867  0.06270 .
Education      1.99828    2.59979   0.769  0.44257
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 161.6 on 393 degrees of freedom
Multiple R-squared:  0.8782,    Adjusted R-squared:  0.8764
F-statistic: 472.5 on 6 and 393 DF,  p-value: < 2.2e-16
```

- 예시: 고객의 특성(X)과 신용카드 잔액(Y)의 관계를 표현한 회귀식

```
> model1 = lm(Balance ~ Income + Limit + Rating + Cards + Age + Education, data = Credit)
> summary(model1)

Call:
lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
    Education, data = Credit)

Residuals:
    Min      1Q  Median
-227.25 -113.15  -42.06

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -477.95809   55.06529  -8.680  < 2e-16 ***
Income        -7.55804    0.38237 -19.766  < 2e-16 ***
Limit          0.12585    0.05304   2.373  0.01813 *
Rating         2.06310    0.79426   2.598  0.00974 **
Cards         11.59156                     0.10174
Age           -0.89240                     0.06270 .
Education      1.99828    2.59979   0.769  0.44257
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 161.6 on 393 degrees of freedom
Multiple R-squared:  0.8782,	Adjusted R-squared:  0.8764
F-statistic: 472.5 on 6 and 393 DF,  p-value: < 2.2e-16
```

①

회귀식

$$Balance = -477.96 - 7.56Income + 0.13Limit$$
$$+2.06Rating + 11.59cards - 0.89Age + 2.00Edu$$

해석

다른 조건들이 동일할 때,

신용등급이 1만큼 증가하면

신용카드 잔액은 평균적으로 2.06만큼 증가

- 예시: 고객의 특성(X)과 신용카드 잔액(Y)의 관계를 표현한 회귀식

```
> model1 = lm(Balance ~ Income + Limit + Rating + Cards + Age + Education, data = Credit)
> summary(model1)

Call:
lm(formula = Balance ~ Income + Limit + Rating + Cards + Age + 
    Education, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max 
-227.25 -113.15  -42.06   45.82  542.97 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -477.95809   55.06529  -8.680  < 2e-16 ***
Income        -7.55804    0.38237 -19.766  < 2e-16 ***
Limit          0.12585    0.05304   2.373  0.01813 *  
Rating         2.06310    0.79426   2.598  0.00974 ** 
Cards         11.59156    7.06670   1.640  0.10174    
Age           -0.89240    0.47808  -1.867  0.06270 .  
Education      1.99828    2.59979   0.769  0.44257    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 161.6 on 393 degrees of freedom
Multiple R-squared:  0.8782,	Adjusted R-squared:  0.8764 
② F-statistic: 472.5 on 6 and 393 DF,  p-value: < 2.2e-16
```

F-test

귀무가설 기각

적합된 회귀식이
통계적으로 유의함

- 예시: 고객의 특성(X)과 신용카드 잔액(Y)의 관계를 표현한 회귀식

```
> model1 = lm(Balance ~ Income + Limit + Rating + Cards + Age + Education, data = Credit)
> summary(model1)

Call:
lm(formula = Balance ~ Income + Limit + Rating + Cards + Age + 
    Education, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max 
-227.25 -113.15  -42.06   45.82  542.97 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -477.95809   55.06529  -8.680  < 2e-16 ***
Income        -7.55804    0.38237 -19.766  < 2e-16 ***
Limit          0.12585    0.05304   2.373  0.01813 *  
Rating         2.06310    0.79426   2.598  0.00974 ** 
Cards         11.59156    7.06670   1.640  0.10174    
Age           -0.89240    0.47808  -1.867  0.06270 .  
Education      1.99828    2.59979   0.769  0.44257    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 161.6 on 393 degrees of freedom
Multiple R-squared:  0.8782,	Adjusted R-squared:  0.8764 
F-statistic: 472.5 on 6 and 393 DF,  p-value: < 2.2e-16
```

③

R^2

약 88%의 설명력

적합된 회귀식이

데이터를 잘 설명

- 예시: 고객의 특성(X)과 신용카드 잔액(Y)의 관계를 표현한 회귀식

```
> model1 = lm(Balance ~ Income + Limit + Rating + Cards + Age + Education, data = Credit)
> summary(model1)

Call:
lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
    Education, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max
-227.25 -113.15  -42.06   45.82  542.97

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -477.95809   55.06529  -8.680  < 2e-16 ***
Income        -7.55804    0.38237 -19.766  < 2e-16 ***
Limit          0.12585    0.05304   2.373  0.01813 *
Rating         2.06310    0.79426   2.598  0.00974 **
Cards         11.59156    7.06670   1.640  0.10174
Age           -0.89240    0.47808  -1.867  0.06270 .
Education      1.99828    2.59979   0.769  0.44257
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 161.6 on 393 degrees of freedom
Multiple R-squared:  0.8782,	Adjusted R-squared:  0.8764
F-statistic: 472.5 on 6 and 393 DF,  p-value: < 2.2e-16
```

④

T-test

Income, Limit, Rating, Age

: 다른 변수들이

적합된 상태에서

설명력을 증가시킴

Cards, Edu

: 다른 변수들이

적합된 상태에서

설명력을 유의미하게 증가X

4

데이터 진단

- **데이터진단, 왜 필요할까?**

일반적인 경향에서 벗어나는 데이터

ex)이상치, 지렛값, 영향점 등

↓

회귀 모형에 큰 영향을 미침

↓

어떻게 해결할까?

표준화 잔차

(Standardized residual)를

이용!

표준화 잔차값 -> 관측치가

경향성에서 벗어나는지 판단

• 표준화 잔차

잔차를 표준화 시켜준 것!

잔차는 y값의 단위에 영향을 많이 받으므로

좀 더 일반화된 상황에서 적용할 수 있게 하기 위해서!

잔 차

$$e = y - \hat{y} = y - X\hat{\beta} = y - X(X^tX)^{-1}X^ty = y - Hy = (I-H)y$$

$$Var(e) = Var((I-H)y) = (I-H)\sigma^2(I-H)^t = \sigma^2(I-H)(I-H)^t = \sigma^2(I-H)$$

$$Var(e_i) = (I - h_{ii})\sigma^2$$

$$\sigma^2(I-H) = \begin{pmatrix} 1-h_{11} & \cdots & -h_{1n} \\ \vdots & \ddots & \vdots \\ -h_{n1} & \cdots & 1-h_{nn} \end{pmatrix} * \sigma^2$$

표준화잔차

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

σ는 모수이므로

알 수 없기 때문에,

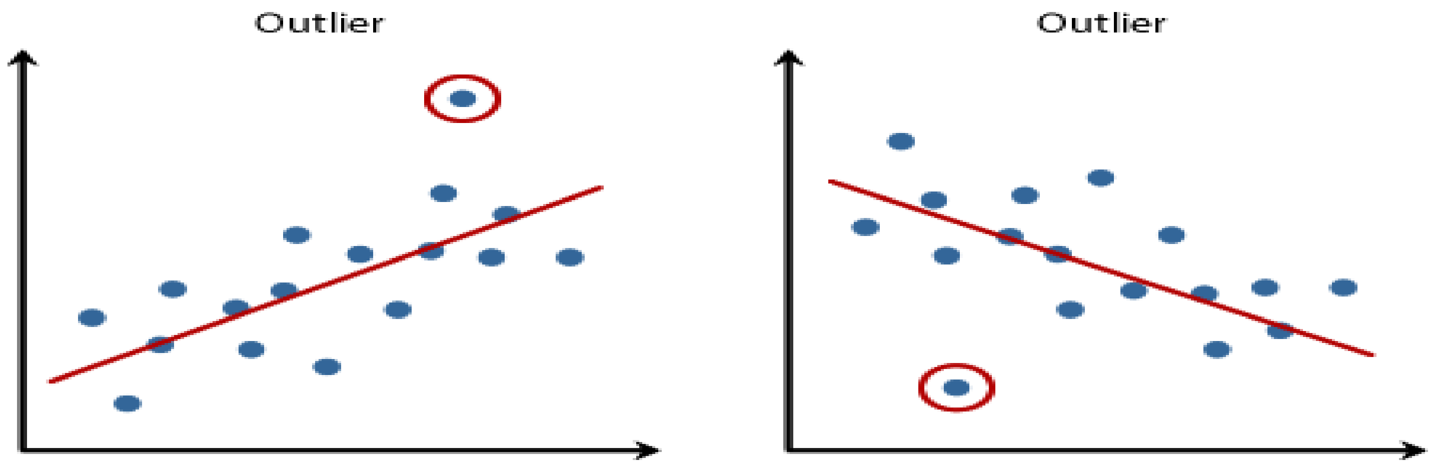
추정량을 넣어준다.

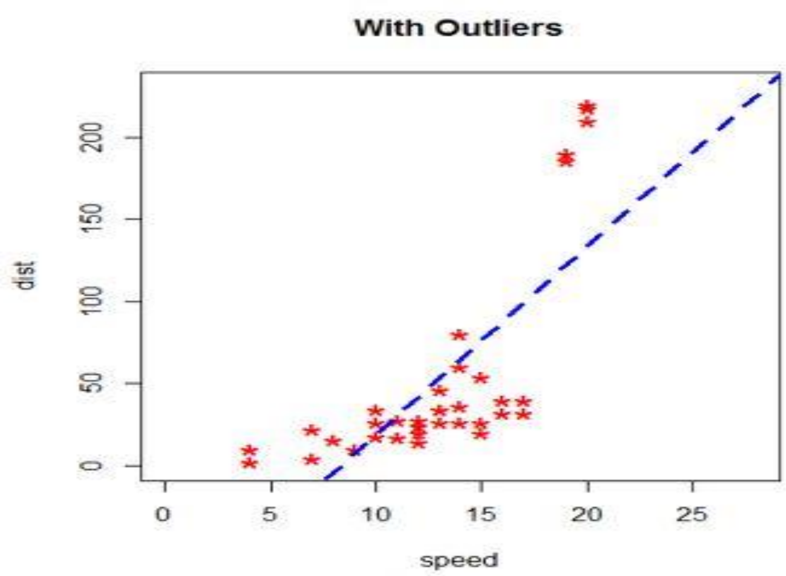
$$\hat{\sigma} = \sqrt{\frac{SSE}{n-p-1}}$$

- # 이상치(Outlier)

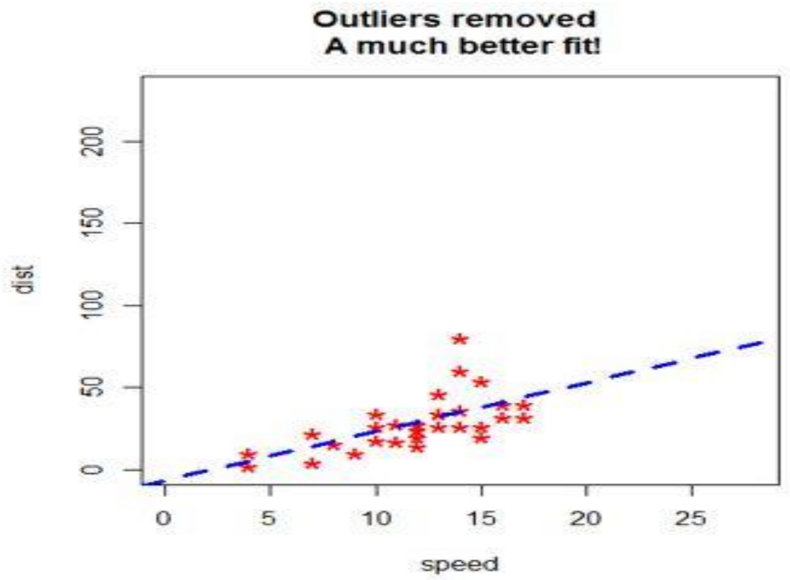
표준화 잔차가 매우 큰 값!

$|r_i| > 3$이면 이상치로 판단!





이상치(Outlier)를 포함한 회귀선

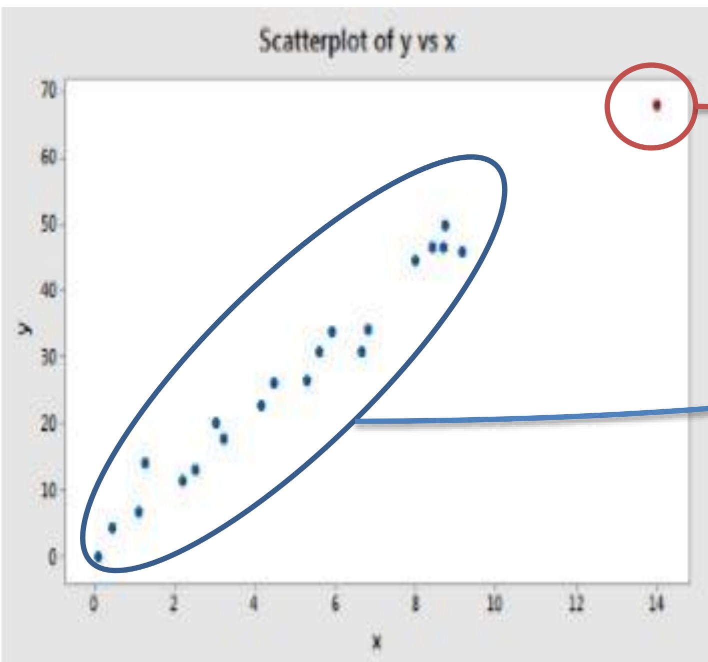


이상치(Outlier)를 포함하지 않은 회귀선

이상치 제거는 모델의
정확성을 높이는데 중요!

- **지렛값(Leverage Point)**

표준화했을 때, x 기준에서 절대값이 큰 값!

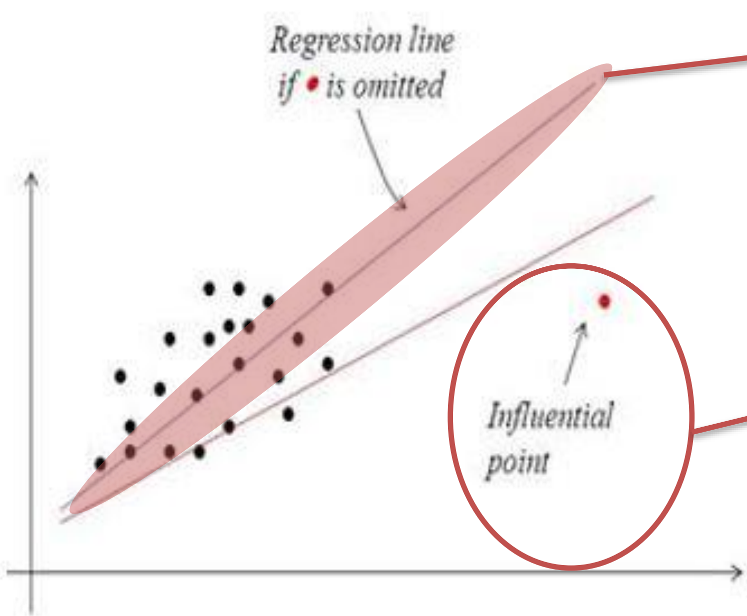


$h_{ii} \geq \frac{2(p+1)}{n}$ 이면 지렛값!

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

- 영향점(Influential Point)

  회귀직선의 기울기에 상당한 영향을 주는 점



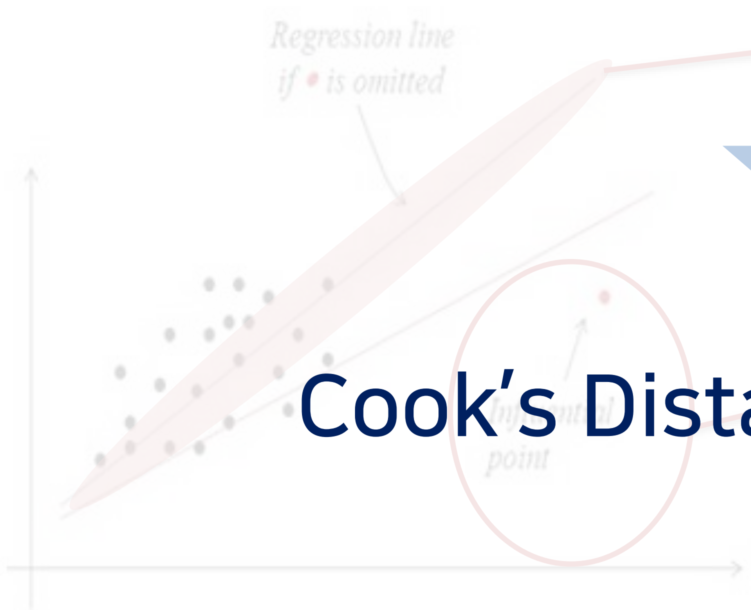
Influential point를 제거한 회귀선

Influential point 하나로 인해
회귀직선의 기울기가 크게 바뀜!

- 영향점(Influential Point)

회귀직선의 기울기에 상당한 영향을 주는 점

영향점은 어떻게 판단할까?



Influential point를 제거한 회귀선

Influential point 하나로 인해 회귀직선의 기울기가 크게 바뀜!

Cook’s Distance를 사용하자!

- Cook’s distance

이상치와 지렛값을 동시에 고려하여, 특정 데이터를 지웠을 때 회귀선이 변하는 정도를 나타내는 지표

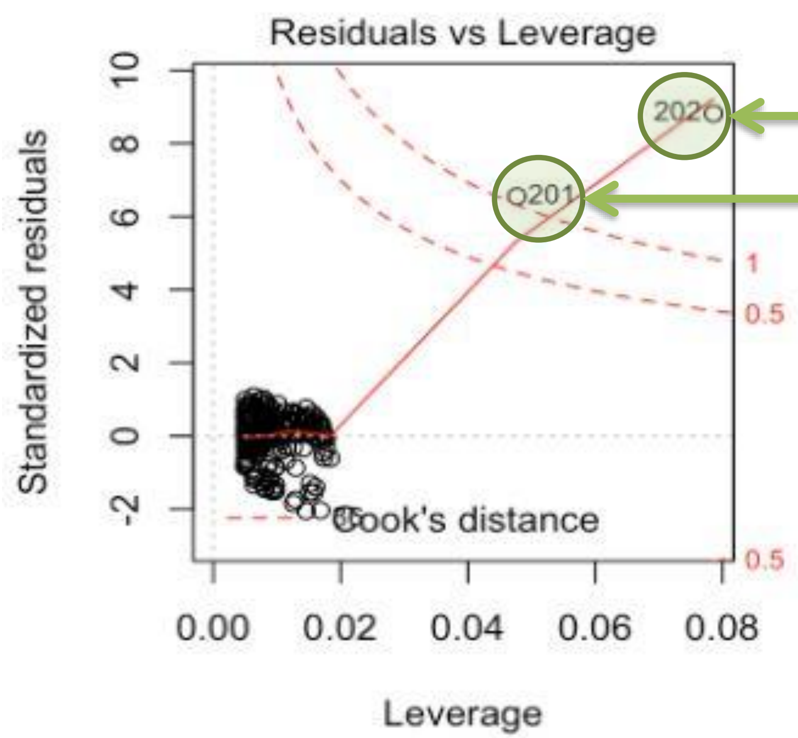
이상치를 고려

지렛값를 고려

$$C_i = \frac{{r_i}^2}{p+1} \times \frac{h_{ii}}{1-h_{ii}}$$

$C_i > 1$ 이면 영향점이라 판단!

- R에서의 예시



Cook's distance를 통해 201, 202번째 데이터값이 영향점(Influential point)임을 알 수 있다!

이러한 영향점들은 어떻게 처리할까?

1) 영향점 삭제

데이터를 삭제할 때에는 항상 신중해야 한다!

왜 영향점이 되었는지 고민해보고 삭제하자!

2) 로버스트(Robust) 모델링

5

로버스트 회귀

- 로버스트(Robust) 회귀란?

  → 건장한, 탄탄한

  이상치의 영향력을 크게 받지 않는 회귀모형

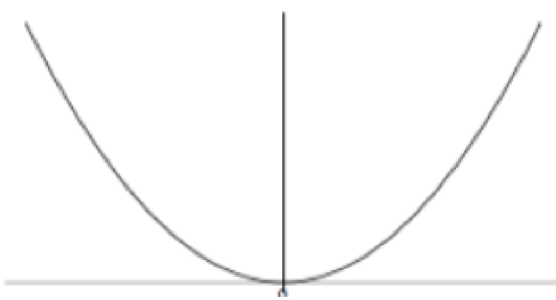
- 로버스트(Robust) 회귀 종류

Median Regression	Huber's M-estimation	Least Trimmed Square

- Median Regression

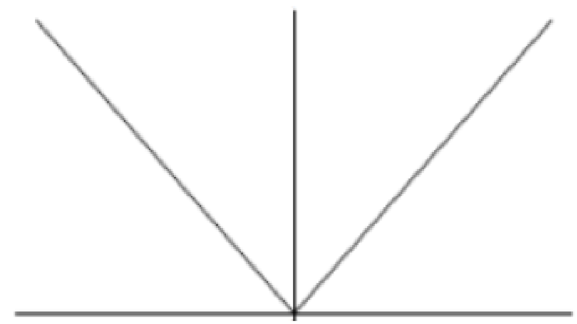
평균보다 중앙값이 이상치의 영향에 덜 받는다는 생각에 기초하여

회귀계수를 추정할 때, X에 따른 Y의 중앙값을 반환하는 모델



최소제곱법은 이상치에 너무

큰 가중치를 둠

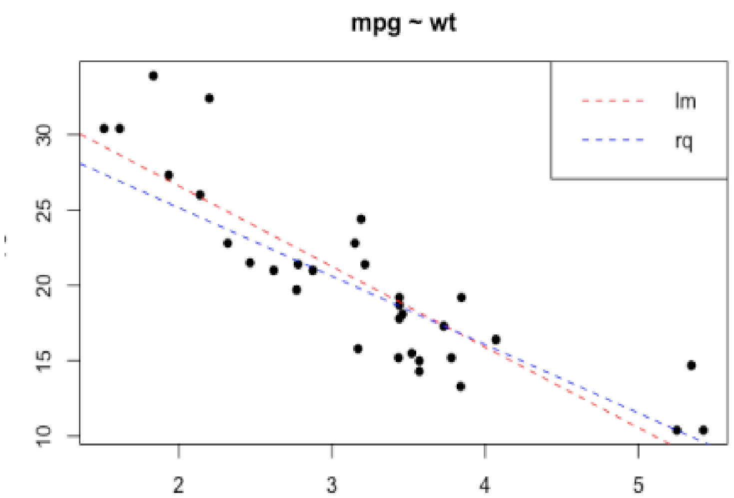


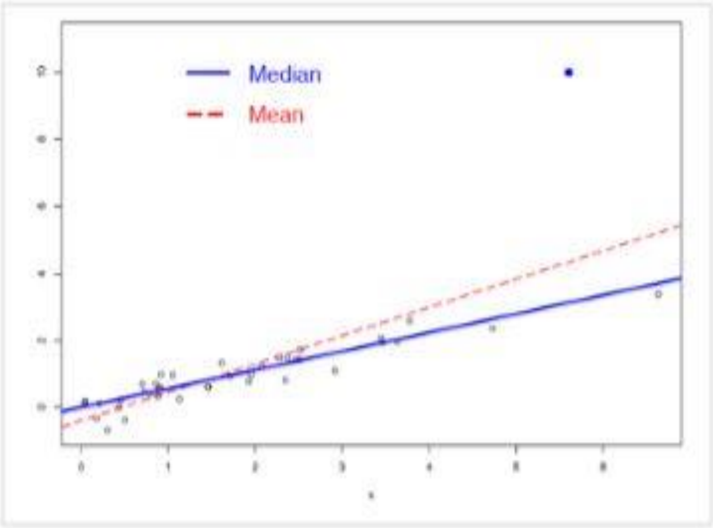
하지만 Median Regressions는

어떠한 경우에도 동일한 가중치

Robust regression: $\underset{\beta}{\operatorname{argmin}} \sum |\varepsilon_i|$

- Median Regression R 예시

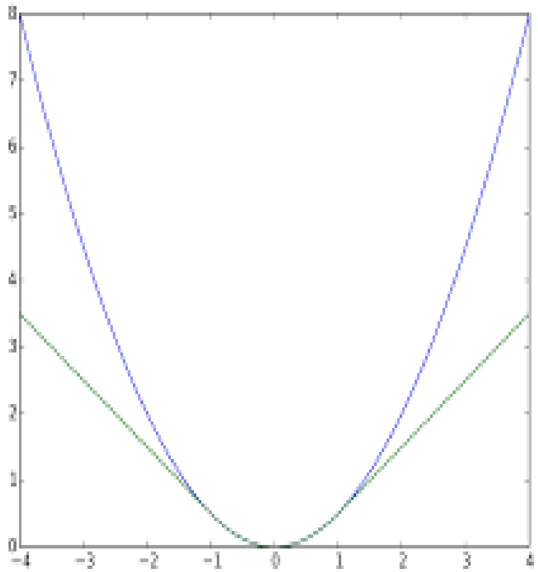




R에서 quantreg 패키지의 rq 함수를 사용하여 표현

- Huber's M-estimation

이상치에 대한 지나친 페널티 부여를 없애는 방식



: LSE 방법

: Huber's M-estimation

특정 잔차(e)가
특정 상수값(c)보다 크면,
패널티를 잔차의 '제곱'이 아닌
1차식으로 바꾸어
이상치에 덜 민감한 회귀계수를 추정

- Huber's M-estimation

이상치에 대한 지나친 페널티 부여를 없애는 방식

LSE 방법

$$\rho(e) = e^2$$

Huber's M-estimation

$$if\ |e| \le c, \quad \rho(e) = \frac{1}{2}e^2$$

$$if\ \ |e| \ge c,\ \rho(e) = c|e| - \frac{1}{2}c^2$$

특정 잔차(e)가

특정 상수값(c)보다 크면,

패널티를 잔차의 '제곱'이 아닌

1차식으로 바꾸어

이상치에 덜 민감한 회귀계수를 추정

- Least Trimmed Square

잔차가 너무 큰 관측치를 제거하고 회귀 계수를 추정하는 방식

$$\hat{\beta} = min \sum_{j=1}^{h} {r_{(j)}}^2 \begin{cases} r_1 \leq r_2 \leq \cdots \leq r_h \\ \frac{n}{2} + 1 \leq h \end{cases}$$

n개의 obs중 h개만 사용하여, $\binom{n}{h}$개의 회귀식 중

가장 잔차제곱의 합이 작은 값을 사용!

단, obs가 별로 없거나 영향점이 존재하지 않는 경우 주의!

2주차 예고

1. 회귀가정
2. 잔차 플랏
3. 회귀 가정 진단과 처방
4. 공간회귀분석

THANK YOU